

# Lessons being learned from an open-source causal AI suite

Emre Kıcıman

[emrek@microsoft.com](mailto:emrek@microsoft.com)

Microsoft Research

with many collaborators

# What I'll cover

- Core challenge: difficulty is causal framing and validity
  - Errors in solving are not (just) algorithmic, but in basic assumptions
- Role of causal software is to scaffold the end-to-end process
  - Reduce burden on practitioner to bridge the gap between domain problems and language of causality
  - Ensure practitioners follow best practices in modeling and validation
- Three fundamental research challenges
  - Eliciting domain knowledge
  - Analysis over unstructured text and image data
  - Expanding how we validate causal analyses

# Causal Tasks in Practice



- What is the impact of a marketing campaign?
- What are the drivers of customer churn?



- How do farmer practices effect soil carbon sequestration?
- Identifying retail demand factors for inventory planning
- Prioritizing maintenance to mitigate climate change effects



- Who doesn't benefit from a drug?
- What are long-term effects of a disease?
- How can AI models generalize across hospitals?

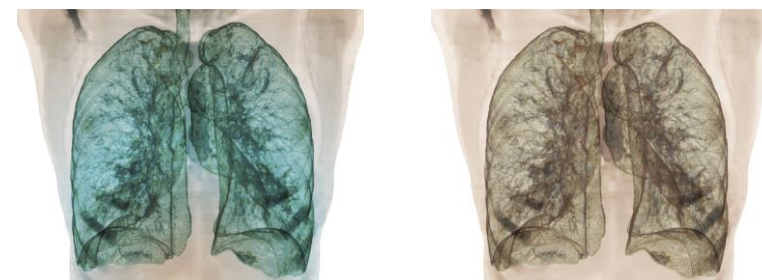
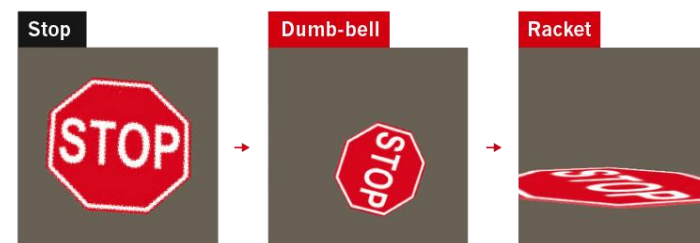
Effect inference, policy optimization, attribution, discovery, robust prediction, ...



Correlational machine learning  
searches for patterns.  
Often finds spurious ones

# ML fails when data patterns change

- Digit recognition accuracy drops to 64% under rotations  
*[Piratla et al. ICML 2020]*
- Object classifier fails when 3D perspective shifts  
*[Alcorn et al. CVPR 2019]*
- 100s of chest scan COVID classifiers identified false correlates (e.g., sitting vs lying down; pediatric scans)  
*[Roberts et al. NMI 2021, Wynants et al. BMJ 2020]*

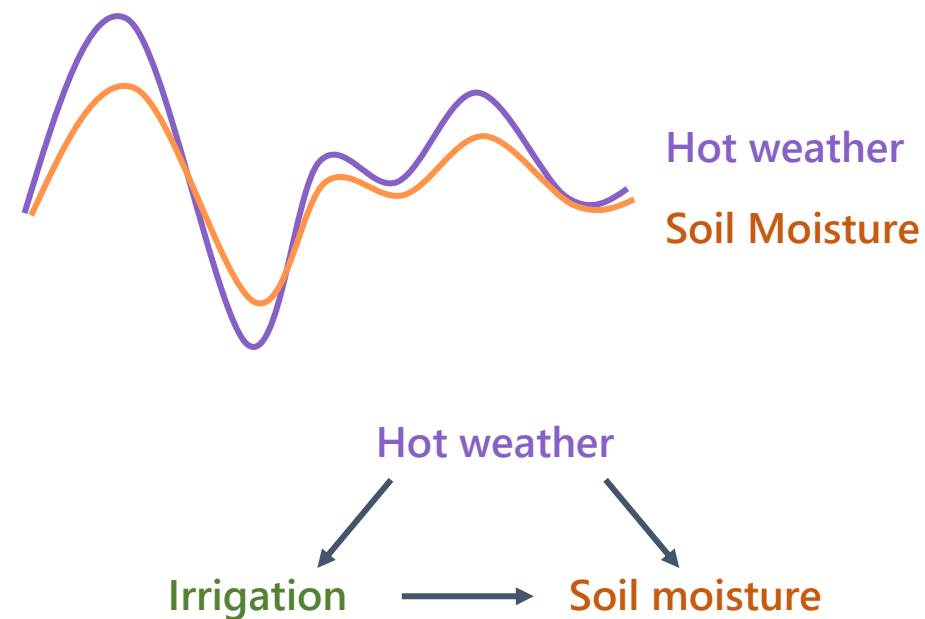


# Active decision-making can change distribution

- Task: Irrigate based on predicted soil moisture levels
- Train models on years of sensor data from real fields
- Q: It's going to be very hot in 2 days. Do I need to water my fields?
- A: No. The model says the soil moisture will be high.

Uh oh! Models only predict high moisture because farmer always watered on hot days!

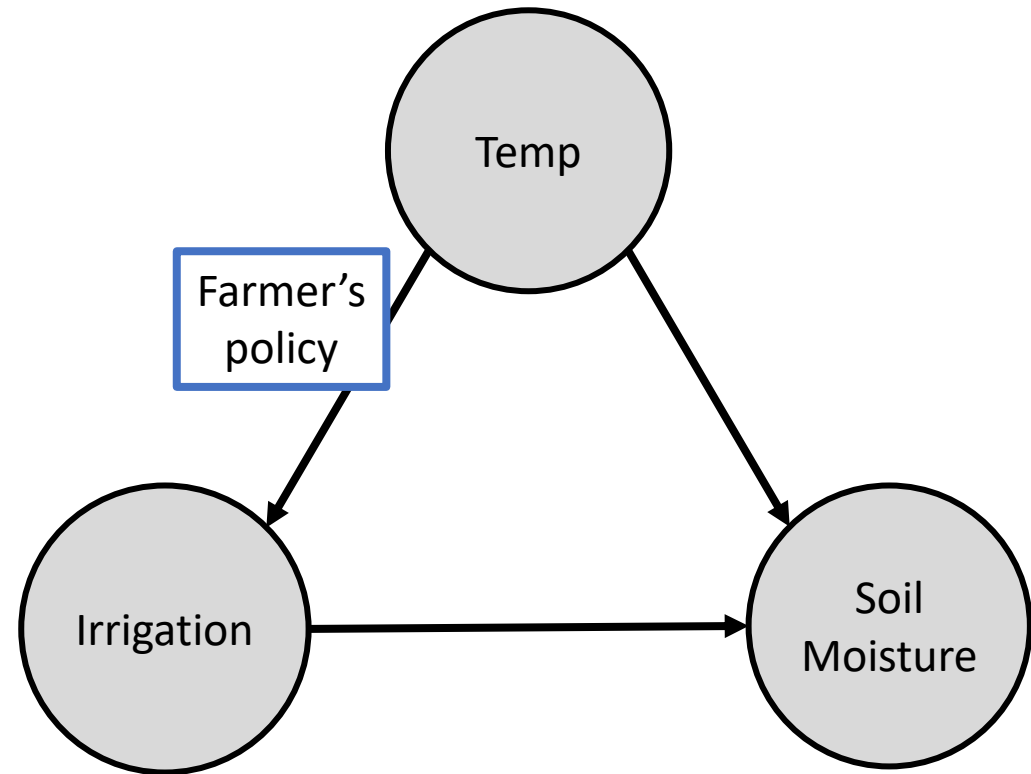
*When we change our action policy based on the ML model, we break the patterns the model learned*



# Example of Spurious Correlations

Have observed data:

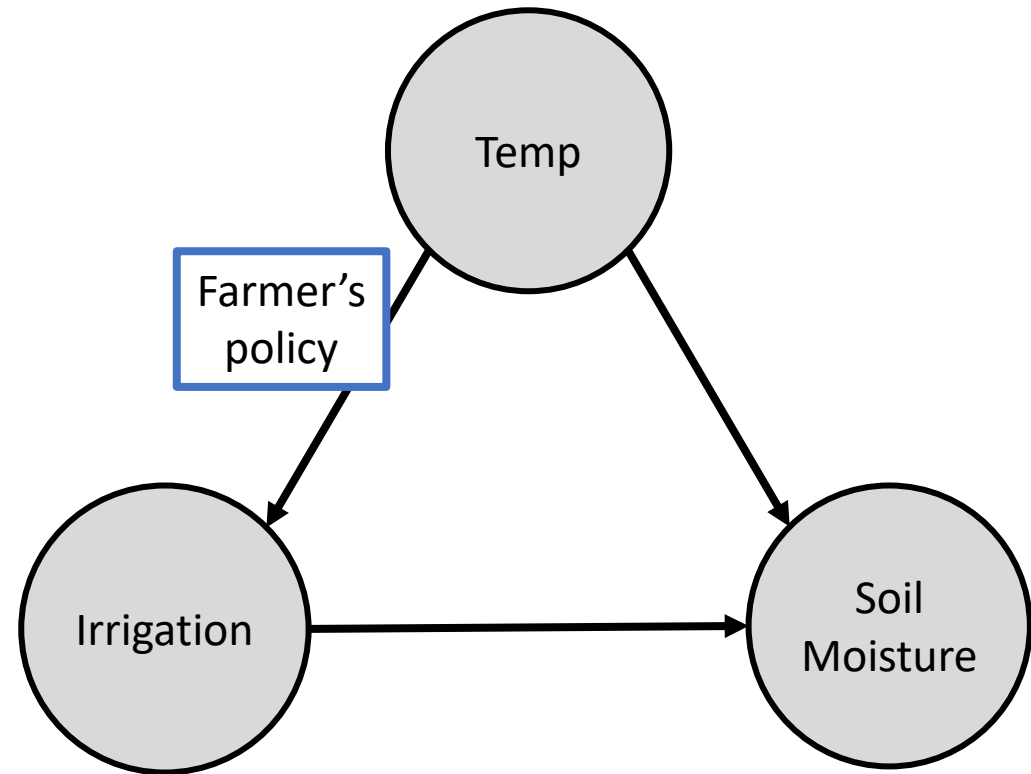
- Temperature is correlated with irrigation action
  - Determined by Farmer's policy
- *Both* predict outcome
  - Underspecified: Multicollinearity



# Example of Spurious Correlations

Have observed data:

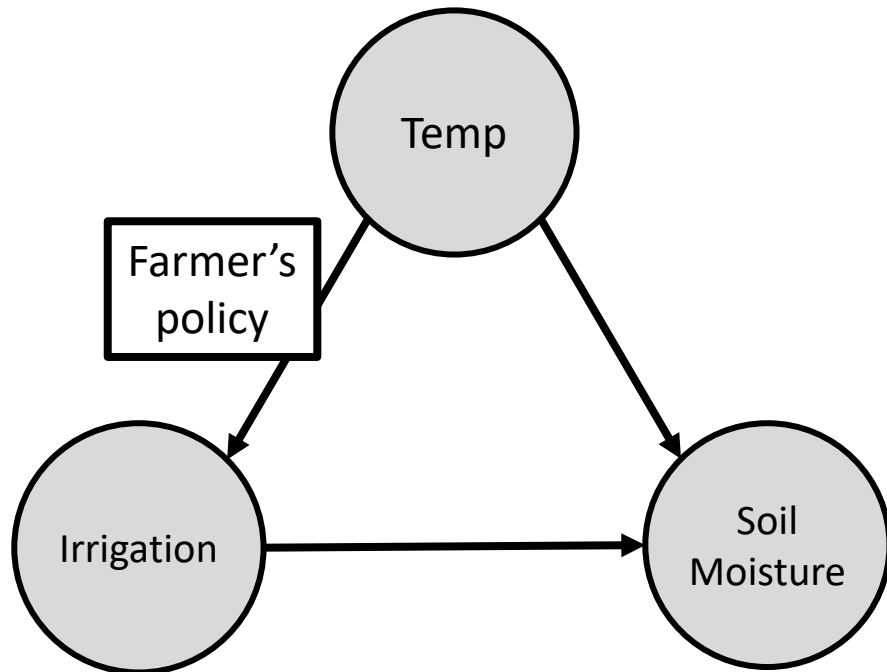
- Temperature is correlated with irrigation action
  - Determined by Farmer's policy
- *Both* predict outcome
  - Underspecified: Multicollinearity
- What if farmer's policy changes?
- How much of outcome caused by temperature vs irrigation?



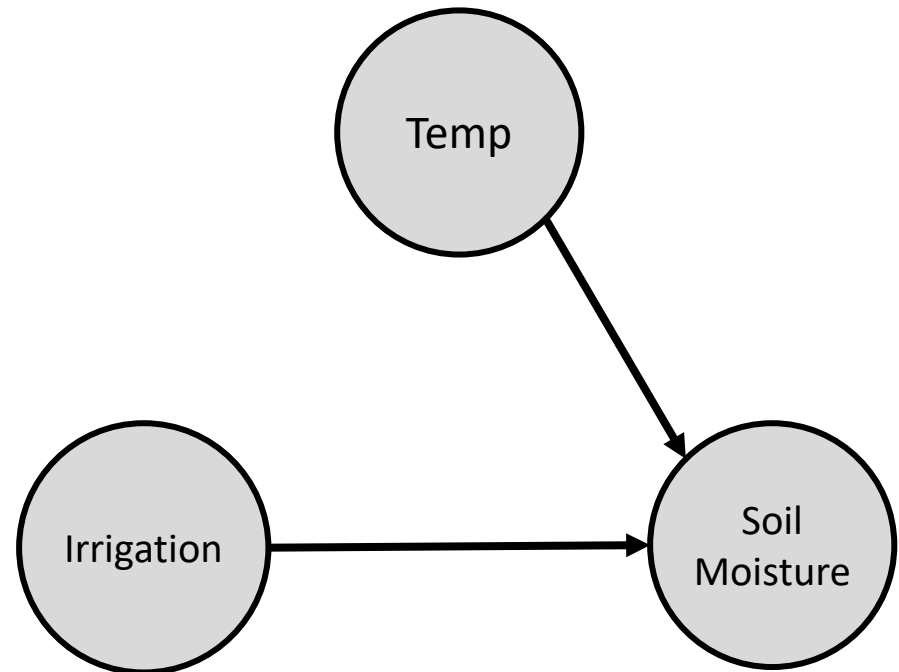


# Disentangling relationships

We observe this joint distribution



We want this joint distribution



# Disentangling relationships

- Causal reasoning enables this
- But relies strongly on initial assumptions
  - domain knowledge, causal structure
- How do we know if we are right?

# Two Fundamental Challenges for Causal Inference

1. Multiple causal mechanisms can be fit to a single data distribution
  - Data alone is not enough for causal inference
  - **Domain knowledge and assumptions come from outside the dataset**
2. We never observe counterfactual world
  - Cannot directly calculate the causal effect
  - Must estimate the counterfactuals
  - **Challenges in validation**



Real World: **do(T=1)**



Counterfactual World: **do(T=0)**

# 1. Measurement and Modeling

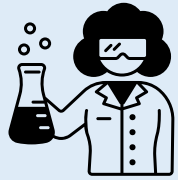
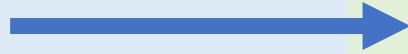
# 2. Identification

# 3. Effect Estimation

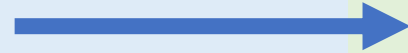
# 4. Validation



data



domain knowledge



causal estimand



causal effect



causal question

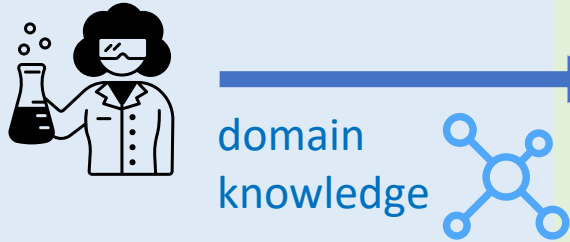


# 1. Measurement and Modeling

# 2. Identification

# 3. Effect Estimation

# 4. Validation



## Construct Validity

Do measurements mean what we think they mean?



# 1. Measurement and Modeling

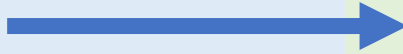
# 2. Identification

# 3. Effect Estimation

# 4. Validation



data

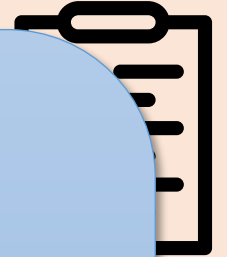


domain  
knowledge



## Causal Assumptions

Capture potential confounders and other causal mechanisms



# 1. Measurement and Modeling

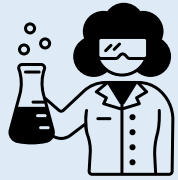
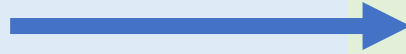
# 2. Identification

# 3. Effect Estimation

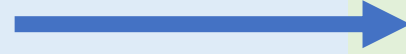
# 4. Validation



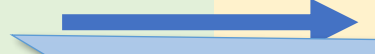
data



domain knowledge



causal estimand



causal question



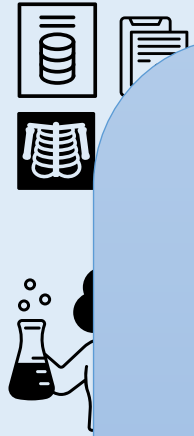
**Analyze assumptions and data to identify valid analysis strategies**

# 1. Measurement and Modeling

# 2. Identification

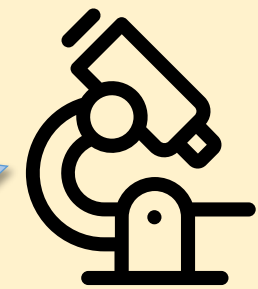
# 3. Effect Estimation

# 4. Validation



**Given a strategy, estimation is a statistical problem**  
Introduces new assumptions on functional forms, etc.

causal and



causal effect →





# 1. Measurement and Modeling

# 2. Identification

# 3. Effect Estimation

# 4. Validation

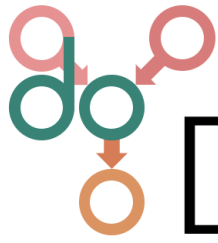


**All assumptions made throughout the analysis must be scrutinized**

Validate, refute, or run sensitivity analysis

al effect





# DoWhy

## Open-source library for causal inference

### Assumptions front-and-center

- Transparent declaration of assumptions
- Evaluation of those assumptions, as possible

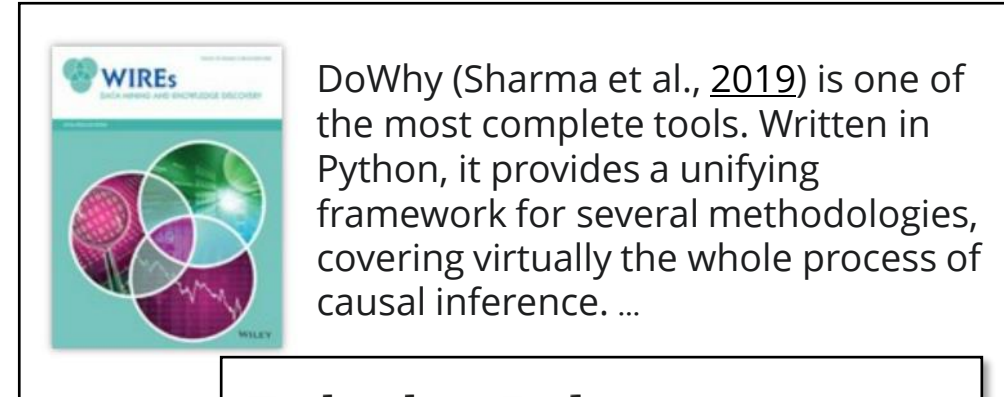
### Most popular causal library on GitHub

- 2M+ downloads, 5.6K stars, 800+ forks)
- Taught in 3<sup>rd</sup>-party tutorials, courses: [O'Reilly](#), [PyData](#), [Northeastern](#), Technion, ...

### Open-source community of 60+ contributors

- MIT, CMU, Johns Hopkins, Columbia; Microsoft, Amazon, and many others

Broad academic and industrial usage, including production deployments



DoWhy (Sharma et al., [2019](#)) is one of the most complete tools. Written in Python, it provides a unifying framework for several methodologies, covering virtually the whole process of causal inference. ...

### Technology Radar

An opinionated guide to technology frontiers

#### 70. DoWhy

**Trial:** Worth pursuing. It's important to understand how to build up this capability. Enterprises can try this technology on a project that can handle the risk.



Judea Pearl  @yudapearl · May 31

Hats off to [@emrek](#) and [@amt\\_shrma](#) for the development of DoWhy and its new PyWhy platform. Next step is for [@MSFTResearch](#) to help starving academia re-educate the thousands ML folks who are stuck on curve fitting and will soon be needed in the workforce.

# Design Principles

- Scaffolding causal analysis process for practitioners
  - Especially causal assumptions
- Separation of causal questions from algorithms and implementations
- Bridge across causal frameworks with common abstractions



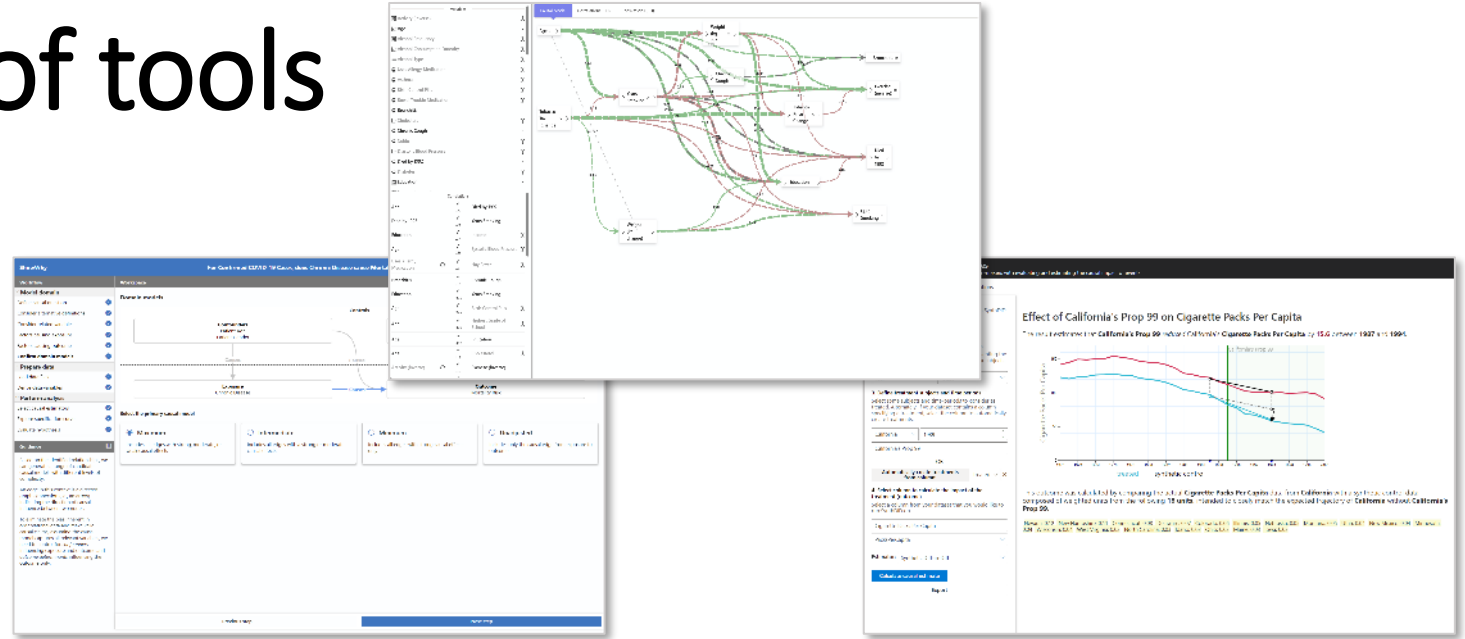
# Grown into a suite of tools

DoWhy

EconML

Causica

ShowWhy



Discovery

Effect Inference  
*(identification & estimation)*

Validation  
&  
Reporting



domain  
knowledge



causal graph



causal effect

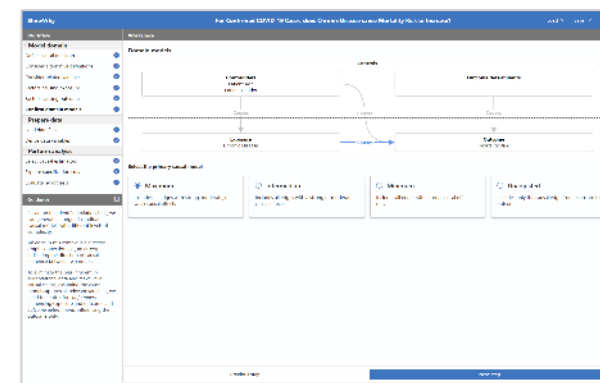
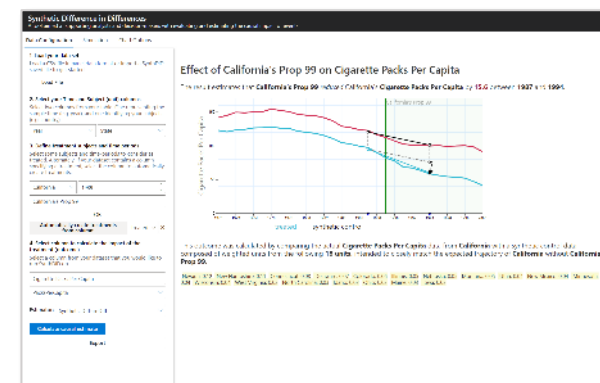
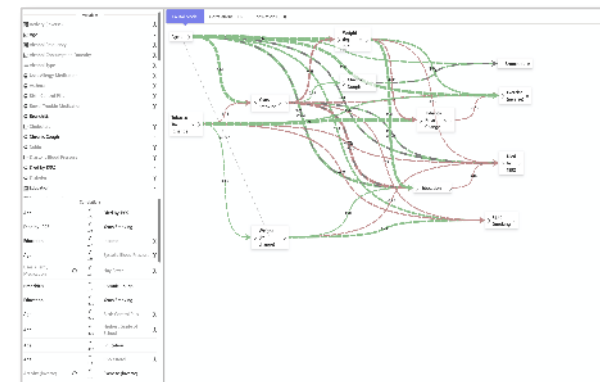


data



# Open source suite for Causal AI

- **DoWhy**: end-to-end library for causal analysis to scaffold best practices
- **EconML**: sophisticated estimation methods based on latest causal ML
- **Causica**: deep learning methods for causal discovery and end-to-end inference
- **ShowWhy**: no-code interactive tools for data analysts



# Community-driven governance at PyWhy

[< Return to Blog Home](#)

## Microsoft Research Blog

### DoWhy evolves to independent PyWhy model to help causal inference grow

Published May 31, 2022

By [Emre Kiciman](#), Senior Principal Researcher; [Amit Sharma](#), Principal Researcher

Share this page [f](#) [t](#) [in](#) [e](#) [s](#)

## Why PyWhy?

PyWhy's mission is to build an open-source ecosystem for causal machine learning that moves forward the state-of-the-art and makes it available to practitioners and researchers. We build and host interoperable libraries, tools, and other resources spanning a variety of causal tasks and applications, connected through a common API on foundational causal operations and a focus on the end-to-end analysis process.

# Applications of causal methods

3 scenarios from open-source usage; and 2 from our deeper partnerships

# Scenario #1: Drivers of key business metrics

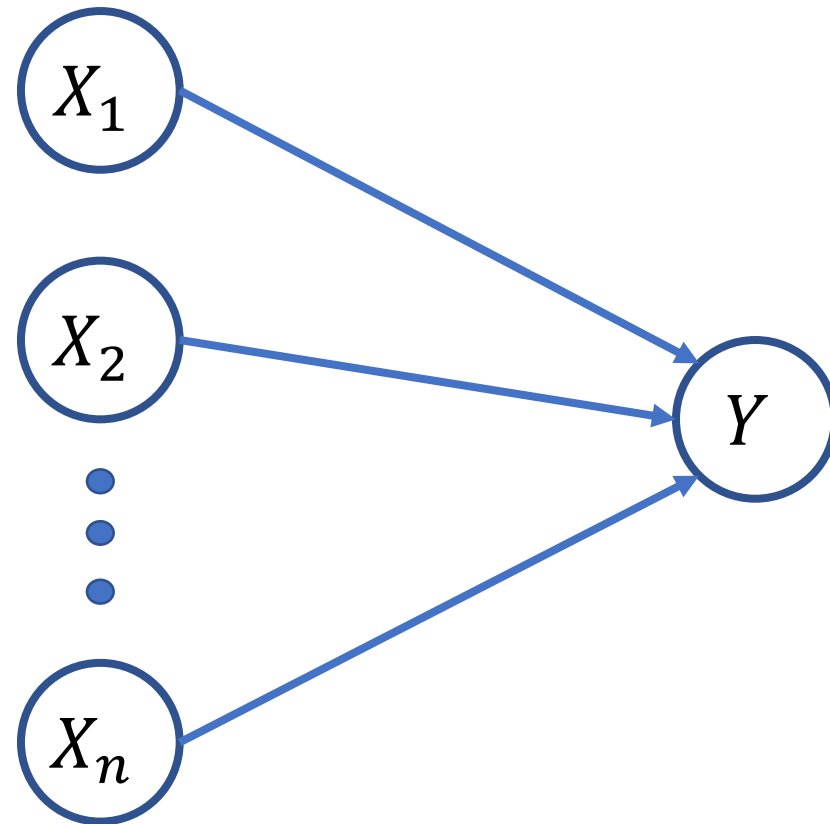
By analyzing the effect of manipulating some controllable input, customers are able to identify the inputs that are most likely to have the greatest impact on business metrics.

For each  $X_i$ :

Calculate  $P(Y|do(X_i))$

Example #1: [Causality example with AirBnB data](#)

Example #2: [Causal Story Behind Hotel Booking Cancellations](#)

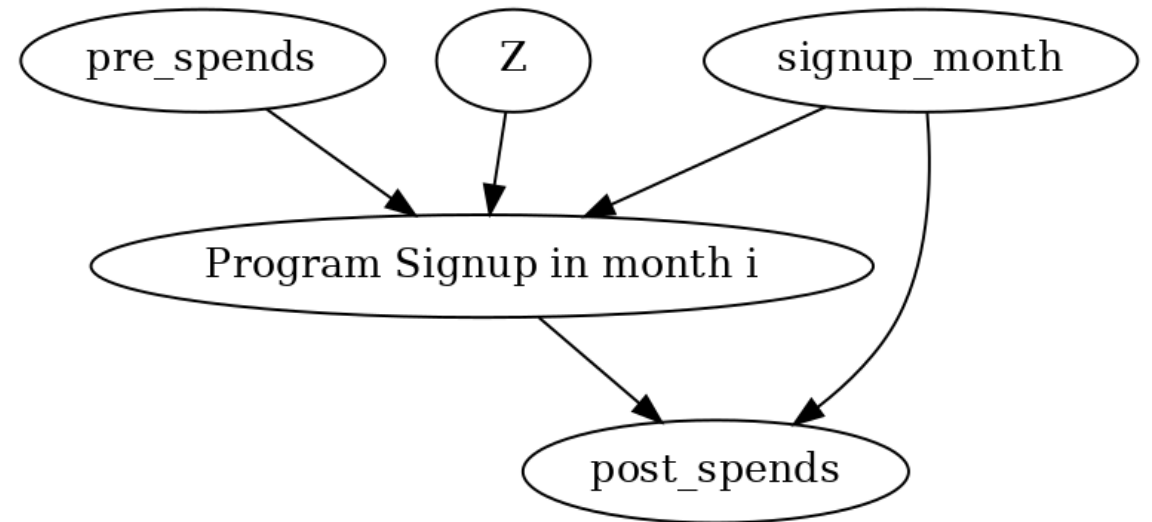




# Scenario #2: Impact of loyalty programs

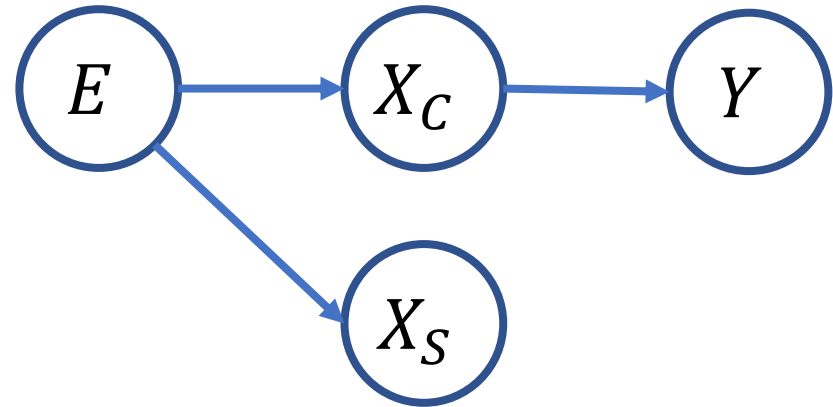
The effectiveness of customer programs is difficult to ascertain, as any increased business from customers participating in such programs is entangled with the customers choice to participate in the program. If we observe that customers who receive loyalty points spend more money, how do we know whether it is the program that is incentivizing their increased spending or if the high-spending customers are choosing to join the program?

Example #1: [Estimating the effect of a Member Rewards program](#)



# Scenario #3: Robust prediction and forecasting

Conventional forecasting models are based on identifying correlations in historical data that have predicted outcomes. Over time, however, these correlations shift and become unreliable. Limiting forecasting models to use only causal features and relationships for prediction is beneficial, especially when the world is changing quickly.



# Deeper partnership – Causal and Bing Ads

## Main driver: A/B experiments come with real cost

- Robust Click Prediction [1]
  - How can we improve off-line click prediction to better estimate effects of policy and system changes in advertising platform?
  - “CTRF” algorithm: combines experimental and observational data for KPI improvement
- Estimating the effect of ad campaigns on user behavior at scale

[1] Zeng et al. “Causal Transfer Random Forest: Combining Logged Data and Randomized Experiments for Robust Prediction,” WSDM 2021

[2] Nabi et al. “Causal Inference in the Presence of Interference in Sponsored Search Advertising”, Frontiers in Big Data 2022.

# Deeper partnership – Soil carbon modeling

- How can we estimate the impact of agricultural practices on soil carbon sequestration at a specific field / farm?
  - Current approaches do not generalize, processes are not well understood
- We are applying causal discovery methods to identify stable processes and improve generalizability of prediction models
- Working with partners to guide large-scale data gathering to improve understanding of processes

# Where we see biggest need for research

## **Frontier #1: People have a hard time encoding domain knowledge**

Research: New sources of domain knowledge and tooling

- Deep causal discovery algorithms [1]
- Algorithms to bootstrap models from trusted simulators

## **Frontier #2: Causal analysis over high-dimensional data, text and images**

Research: Deep Causal Representation Learning

- Disentangle high-dimensional data to match semantics of cause-effect mechanisms [2]
- Apply DNNs for non-parametric and high-dimensional modeling throughout pipeline

## **Frontier #3: New approaches to validation without ground-truth labels**

- Empirical validation of methods with A/B experiments when feasible
- Refutations, sensitivity analyses, and analytical tests of core assumptions [3]

 [1] Deep End-to-End Causal Inference, Geffner et al. [arxiv:2202.02195](https://arxiv.org/abs/2202.02195)

 [2] Modeling the data-generating process is necessary for out-of-distribution-generalization, Kaur et al. [arxiv:2206.07837](https://arxiv.org/abs/2206.07837)

 [3] Long story short: Omitted Variable Bias in Causal Machine Learning, Chernozhukov et al. [arxiv:2112.13398](https://arxiv.org/abs/2112.13398)

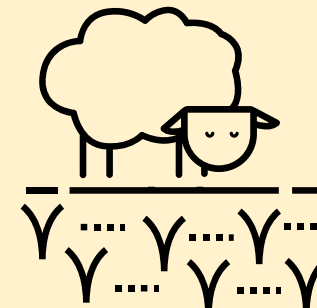
# #1 Domain knowledge and tooling

- Causal ML needs data *and* domain knowledge
  - Formal reasoning over domain knowledge
- Eliciting domain knowledge is hard
  - Not enough experts
  - System not always understood
- New approaches to elicit domain knowledge
  - Domain-specific reusable templates
  - Bootstrap from simulators
  - Beyond DAGs: dynamic mechanisms and more
  - Can we use LLMs as a store of domain knowledge?

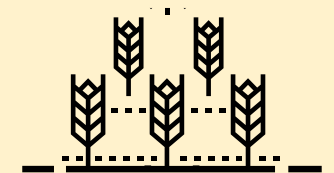


Soil Organic Matter  $\xrightarrow{?}$  Dissolved Oxygen

No correlation



Related

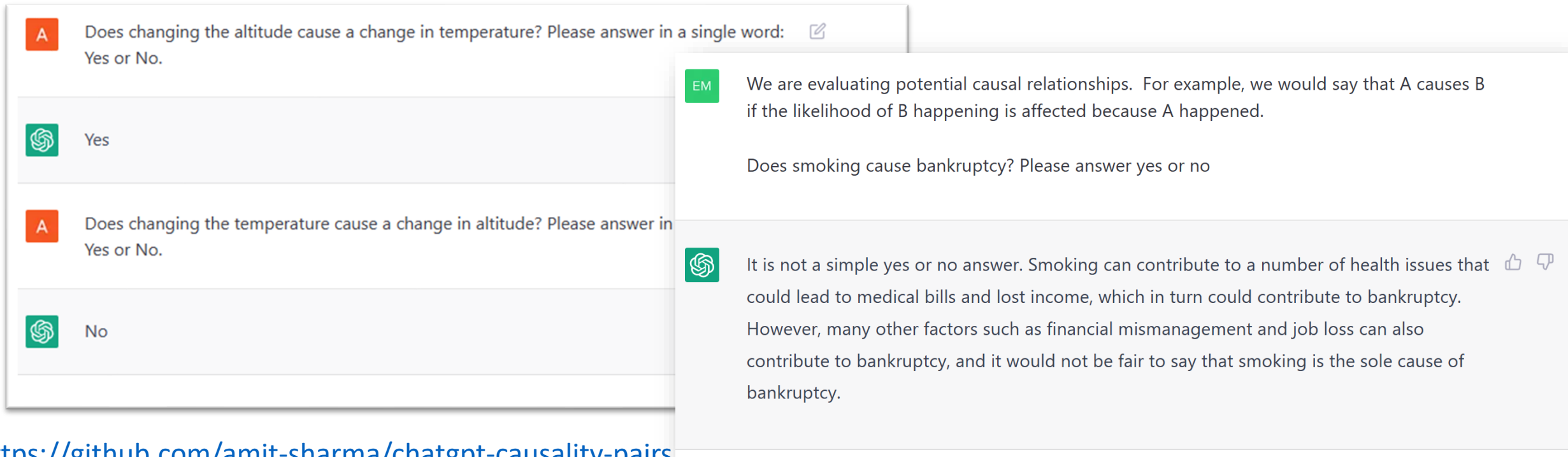


Domain Knowledge: Saturation of Soil Organic Matter in some datasets occludes effects


# Bootstrap causal graphs with LLMs?

ChatGPT obtains an accuracy of 92.5%. In comparison, the best known accuracy using conventional discovery methods is 70-80% [[Mooij et al. \(2016\)](#), [Tagasovska et al. \(2020\)](#), [Compton et al. \(2020\)](#), [Salem et al. \(2022\)](#)].


But does not know how to analyze data to find new relationships and estimate new effects



**A** Does changing the altitude cause a change in temperature? Please answer in a single word: Yes or No.


 Yes

**A** Does changing the temperature cause a change in altitude? Please answer in Yes or No.

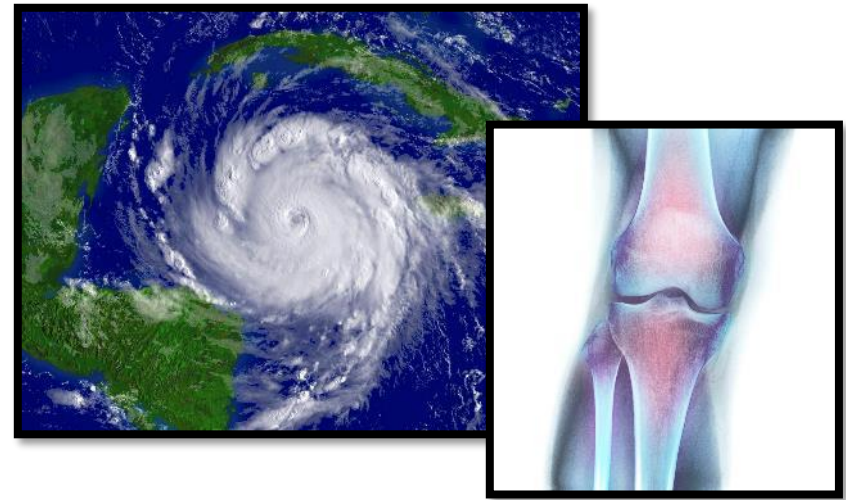
 No

**EM** We are evaluating potential causal relationships. For example, we would say that A causes B if the likelihood of B happening is affected because A happened.

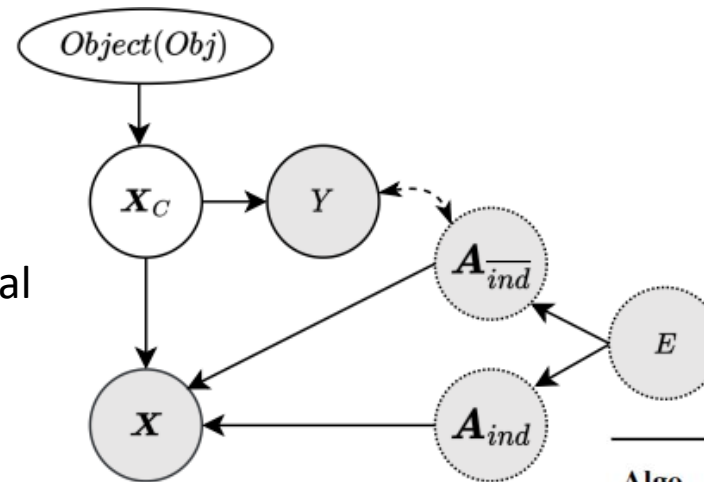
Does smoking cause bankruptcy? Please answer yes or no

 It is not a simple yes or no answer. Smoking can contribute to a number of health issues that could lead to medical bills and lost income, which in turn could contribute to bankruptcy. However, many other factors such as financial mismanagement and job loss can also contribute to bankruptcy, and it would not be fair to say that smoking is the sole cause of bankruptcy.

# #2: Causal analysis over high-dimensional text and images



- Causality beyond tabular data
  - Text and images do not directly map to nodes in a graph
- Deep causal representation learning
  - Disentangle semantic elements
  - Loss function in deep learning based on prior causal knowledge
- Learned embedding useful for robust prediction, effect inference
  - *CACM* algo (Kaur et al.) achieves state-of-the-art results



Algo.	Accuracy		
	color	rotation	col+rot
ERM	30.9 ± 1.6	61.9 ± 0.5	25.2 ± 1.3
IRM	50.0 ± 0.1	61.2 ± 0.3	39.6 ± 6.7
VREx	30.3 ± 1.6	62.1 ± 0.4	23.3 ± 0.4
MMD	29.7 ± 1.8	62.2 ± 0.5	24.1 ± 0.6
CORAL	28.5 ± 0.8	<b>62.5 ± 0.7</b>	23.5 ± 1.1
DANN	20.7 ± 0.8	61.9 ± 0.7	32.0 ± 7.8
C-MMD	29.4 ± 0.2	62.3 ± 0.4	32.2 ± 7.0
CDANN	30.8 ± 8.0	61.8 ± 0.2	32.2 ± 7.0
<i>CACM</i>	<b>70.4 ± 0.5</b>	62.4 ± 0.4	<b>54.1 ± 1.3</b>

[5] Peyrard, Ghotra, Josifoski, Agarwal, Patra, Carignan, Kıcıman, Tiwary, West, Invariant Language Modeling (EMNLP 2022)

[6] Kaur, Kıcıman, Sharma. Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization (ICLR 2023 – Notable Top 25%)



# #3: Expanding Approaches to Validation

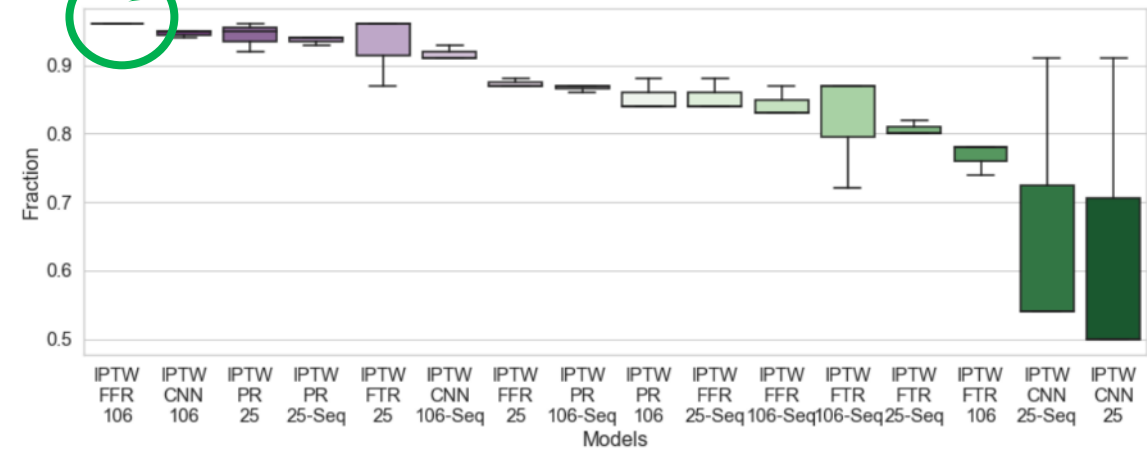


Real World:  $do(T=1)$

Counterfactual:  $do(T=0)$

- ML cross-validation uses ground-truth labels
  - But “what-if” answers are not observable
- Test core assumptions instead
  - Modeling, algorithmic, statistical
  - Validations, refutations, sensitivity analyses
- New targets for scaffolding at scale
  - Small A/B experiments; observational monitoring of new decision policies?

**Empirical results:**  
Models and data design with tightest sensitivity bounds gave best causal results (A/B tested)



[7] Chernozukhov, Cinelli, Newey, [Sharma, Syrgkanis](#), Long Story Short: Omitted Variable Bias in Causal Machine Learning, [NBER working paper](#)

[8] [Xu, Mahajan, Manrao, Sharma, Kiciman](#), Split-Treatment Analysis to Rank Heterogeneous Causal Effects for Prospective Interventions, [WSDM 2020](#)

[9] [Sharma, Syrgkanis, Zhang, Kiciman](#). DoWhy: Addressing Challenges in Expressing and Validating Causal Assumptions, [ICML Workshop](#)

# Questions?

✉ emrek@microsoft.com 🗨 @emrek@hci.social 🐦 @emrek

Links:

- Causica: <https://github.com/microsoft/causica>
- DoWhy/PyWhy: <https://pywhy.org/>
- EconML: <https://github.com/microsoft/econml>
- ShowWhy: <https://github.com/microsoft/showwhy>

Thanks to the whole team:

- Eleanor Dillon, Darren Edge, Adam Foster, Agrin Hilmkil, Joel Jennings, Chao Ma, Robert Ness, Nick Pawlowski, Amit Sharma, Cheng Zhang, Keith Battocchi, Mónica Carvajal, Denise Chen, Nathan Evans, Jonathan Larson, Andrés Morales Esquivel, Friederike Niedtner, Dayenne de Souza, Christopher Trevino, Ha Trinh, Fabio Vera, Robert King, Ahmed Mostafa, and Thomas Kapler
- Our collaborators in the PyWhy organization, Peter Götz, Patrick Blöbaum, Kailash Budhathoki, and all our open-source contributors